我国图书情报学作者自引行为研究初探*

李昕雨　雷佳琪　步一
北京大学信息管理系　北京 100871

摘　要：[目的/意义] 作者自引是学术界常见的现象，是引文分析和科学评价中考虑的重要因素，但如何在科学评价中处理自引一直存在争议。探索作者自引规律能够为适当处理作者自引提供建议。[方法/过程] 首先对作者自引相关的研究成果作了梳理，然后使用中国图书情报学领域期刊的数据，以及回归分析等统计学方法在作者和文献层次对作者自引情况进行分析。[结果/结论] 实证研究表明，中国图书情报学领域的作者自引行为总体上较少，作者自引频次与其发表文献数量有强正相关性，作者自引对h指数的影响比较小，文献自被引对其被他引频次有正向影响等。

关键词：作者自引　引文分析　图书情报学　科学评价
分类号：G253.1

1　引言

自引，是学术界一种常见的引文现象。对作者、期刊、机构、国家等不同的主体来说，自引的定义有所不同[1-2]。其中，期刊自引、作者自引两种自引现象一般是自引分析的重心。引文分析是文献计量学研究的重要方法，自引分析是引文分析的重要组成部分。自引行为能够在一定程度上揭示主体引用的意图、主体研究的连续性等，因此自引研究对于文献计量学以及科学评价来说至关重要。目前，科学评价中常用的传统指标，如期刊影响因子、h指数等指标以被引次数作为基础，而在使用被引数这一指标时，多数情况下是将自引计算在内的，并不区分自引和他引，因此主体有可能通过操纵自引来提高其科学评价。此外，为了使科学评价更加客观有效，作为科学评价数据库之一的期刊引证报告（Journal Citation Report，JCR）开始提供一种不含自引的科学评价指标，即他引影响因子。

自引在学术界是一个饱受争议的话题。有时，期刊的自引已成为其追求高被引和高影响因子的操纵手段之一。JCR每年都会发布“镇压”的期刊名单，不再为这些期刊提供索引服务。2021年JCR发布10种被镇压的期刊，主要原因是过度自引等异常引用行为导致其期刊影响因子和排名有所失真[3]。因此，一些学者认为不当自引行为已极大地影响了科学评价[4]。然而，也有一些学者为自引“正名”，认为自引大多属于正常引用行为[5]。M. Schreiber归纳了自引产生的3个原因：①研究需要使用先前的实验设置、理论模型、结果和结论，但为了避免重复叙述进行了自引，这种自引是合理的；②由于每个人都最了解自己先前的论文，因此在后续研究时引用自己先前的论文相对容易，这种自引有一定争议；③自引是为了提高个人的被引频次、h指数等，这种自引属于无可争议的不当自引[6]。

E. Garfield认为自引本无所谓好坏[7]，关键在于如何看待和利用它。当前状况下，寻找能够区分过度自引和适度自引的方式，探索在科学评价中适当地处理自引的方法，是解决自引争议问题的方向所在。本文将从以下两个方面展开分析：①总结归纳与自引相关的重要研究；②以我国图书情报学领域中文文献数据为例，以作者自引为研究对象，分析与作者自引的相关指标具有相关性的文献计量学因素并探讨中文图书情报学领域作者自引对科学评价的影响程度。

2　研究现状

自引在学术界是普遍的。E. Garfield指出，仅以第

* 本文系国家社会科学基金一般项目“基于全文本分析的数据科学范式及其演化研究”（项目编号：20BTQ054）研究成果之一。
作者简介：李昕雨，硕士研究生；雷佳琪，本科生；步一，助理教授，研究员，博士，通信作者，E-mail：buyi@pku.edu.cn。
收稿日期：2022－07－29　修回日期：2022－09－25　本文起止页码：162－171　本文责任编辑：徐健

一作者计算，自引在全部引文中所占比例约为 10%，若全部合著者都计算在内该比例会更大[8]。同时，自引对科学评价也有重要的影响，因此如何处理自引是科学评价中亟待解决的问题。

2.1 自引测量指标的定义

在自引相关研究中，“自引率”这一定义的使用存在混淆[9]。更准确的做法是针对于“自引证率”和“自被引率”作区分。E. Garfield 在 1975 年发布的 JCR 前言中阐述：期刊的自引率（self-citation rate）有两个，一个是自引证率（self-citing rate），即自引引文在全部参考文献中的比例，另一个是自被引率（self-cited rate），即自引引文在全部被引频次中的比例[10]。自引证率属于自引的共时性指标，而自被引率则是历时性指标。从广义上讲，自引证率和自被引率的计算公式如下：

$$\text{自引证率}=\frac{\text{某主体引证自身的数量}}{\text{某主体引证总数}} \quad \text{公式（1）}$$

$$\text{自被引率}=\frac{\text{某主体被自身引用的数量}}{\text{某主体被引总数}} \quad \text{公式(2)}$$

过去的研究中，对于作者自引，学界没有一个统一规范的定义。由于一篇文献可能由多个作者所著，并且作者有署名顺序的差异，因此对于作者自引可能出现多种定义。目前，多数研究将作者自引定义为一位作者引用了他自身发表的文献。这也是本文所采用的定义，在此定义下作者自引证率和自被引率计算公式如下：

$$\text{作者自引证率}=\frac{\text{作者引证自己文献的数量}}{\text{作者引证总数}} \quad \text{公式（3）}$$

$$\text{作者自被引率}=\frac{\text{作者被自身引用的数量}}{\text{作者被引总数}} \quad \text{公式(4)}$$

值得注意的是，本文使用的作者自引的定义是考虑全作者的作者自引，并不要求自引作者是否为施引论文或被引论文的第一作者。但是，也有少数研究由于数据限制等原因，将两篇有引用关系的文献第一作者为同一人的现象视为作者自引[11－12]。作者自引证率能够反映自引在知识来源中的重要程度，而自被引率能反映自引在学术影响力中的重要程度。

2.2 自引的相关因素

经过文献调研，本文发现作者自引和作者自被引行为与以下因素相关：

（1）文献出版后经过的时间：文献出版后，划定时间窗口越长，作者自被引数在作者所有被引中的份额越低[13－15]。

（2）学科：不同学科间作者自引频率有差异，例如，H. Snyder 和 S. Bonzi 的研究发现物理学科的作者自引率约为 15%，而社会科学和人文科学的作者自引率分别为 6% 和 3%[16]。

（3）性别：一些研究发现，男性作者自引频率高于女性，但是这种差异可能源于男女生产力上的差异[17－18]。

（4）作者生产力/期刊论文数量：研究发现，作者生产力与作者自引量成正相关[14]，出版量明显增加的期刊自被引率比较高[19]。

（5）合著情况：M. R. Davarpanah 和 F. Amel 发现，对于一篇文献，其作者人数和文献的作者自引证数成正相关[14]。蒋颖等发现文献第一作者的自引百分比都远高于其他作者，各作者自引证数一般按照顺序递减[20]。

（6）被引数：D. W. Aksnes 发现，作者总被引数和自被引率呈现负相关关系，总被引数越大，作者自被引率越低[15]。

作者自引与上述因素存在的相关关系还需要更多的研究来验证，而与作者自引相关的其他因素还有待被发现。

2.3 自引对科学评价的影响

自引现象对期刊影响因子、学者 h 指数、论文被他引数量都可能造成影响。过去有许多学者研究了自引对科学评价指标的影响，但由于数据、方法差异较大，得出的结论不尽相同。

期刊自引可能影响期刊影响因子。一些学者认为自引对期刊影响因子有正面影响[21－22]。例如，A. Fassoulaki 等以麻醉学领域六种期刊中的文献为例，证明自引证率和期刊影响因子有显著正相关性[22]。然而，一些研究发现自引可能对期刊影响因子没有显著影响，甚至可能有负面影响[23－26]。例如，J. M. Campanario 和 A. Molina 在 1998－2006 年 JCR 中找到 123 种影响因子连续四年下降的期刊，发现自引行为并不能使他们的影响因子明显增加，因此认为不能通过自引行为操纵影响因子[26]。还有一些研究认为自引对期刊影响因子的影响与期刊影响力或年度等因素有关[27－29]。

作者自引可能对文献影响力有影响，有研究表明，自引能够增加被他引的数量。J. H. Fowler 等使用

挪威的引文数据，通过泊松回归方法发现，作者每增加一次自引后，到第二年他引将会增加大约1次，到第四年大约累计增加2.83次，到第10年累计增加3.65次[30]。

另外，作者自引也可能对h指数及其排名造成影响。学者们对自引对h指数的影响也持不同观点。部分学者认为，自引对h指数的影响比较小[31-32]，例如查颖以《图书馆、情报与文献学学术影响力研究报告(2000-2004)》中论文被引次数排名前10位的学者为调研对象，发现是否剔除自引对h指数及其排名的影响较小，剔除自引后h指数平均下降5.6%[31]。但也有研究认为自引会对h指数产生较大影响，如M. Schreiber经统计认为自引会对h指数产生较大影响，特别是在对h指数较低的青年学者进行科学评价时应当剔除自引[6]。

基于以上研究，本文将针对我国图书情报学领域的自引现象进行实证研究。在本文的第四部分，将对自引与时间、生产力等因素的关系，文献的作者自引对他引数量的影响，自引对h指数的影响等重要问题进行研究。

3 数据与方法

从中国知网数据库获取中国图书情报学领域17种期刊发布于1955-2017年间的98 948篇文献的书目数据，及其参考文献(仅含知网所收录的参考文献)和截至2018年的引证文献数据，用于本文研究。这17种期刊是通过第七版《中文核心期刊要目总览》[33]，与2017-2018年版CSSCI来源期刊目录交集所确定的图书馆学情报学领域的期刊，涵盖了图书情报学领域大多数比较重要的期刊和文献。数据清洗过程中，删除了出版年份信息错误或缺失、作者信息缺失以及参考文献数为0的25 782篇文献，剩余发表于1957-2017年间的73 160篇文献，以及这些文献的38 556名作者。表1为数据集中各期刊文献数量。其中，发表文献最多的是《图书情报工作》。图1显示了文献发表时间分布，由于1957-1990年文献数量较少，在图1左侧另做展示。

本文的研究目标主要是以情报学领域为例，通过探索自引相关指标的相关因素以及自引与自被引行为对论文影响力的影响，来对自引这一现象有更透彻的理解。本文提出以下研究问题：①我国图书情报学的

表1 数据集中各期刊刊载文献数量

期刊名称	文献数量/篇
图书情报工作	10 150
现代情报	9 634
情报科学	6 586
图书馆论坛	5 395
图书馆建设	4 647
情报理论与实践	4 580
图书馆工作与研究	4 510
图书馆理论与实践	4 419
情报杂志	3 817
现代图书情报技术	3 069
情报资料工作	2 739
情报学报	2 618
图书馆杂志	2 369
图书情报知识	2 363
大学图书馆学报	2 327
图书馆	1 990
中国图书馆学报	1 947
合计	73 160

注:《现代图书情报技术》于2017年更名为《数据分析与知识发现》

自引水平如何？自引证率和自被引率有什么特征？②自引的频次和比率与文献出版后时间、作者生产力、作者合作情况是否有相关性？有怎样的相关性？③作者自引对作者的h指数及其排名有怎样的影响？④文献中的作者自引是否会影响文献被他引的频次？

基于以上的研究问题，本研究的数据处理与分析主要通过Python、SPSS、Stata等工具进行。在研究自引对文献影响力的影响时，本文采用负二项回归模型。当因变量为计数变量，即事件发生的数目时，应考虑使用计数模型，这种模型适用于因变量是离散的整数且数值小、取零的个数较多的情况。计数模型包括泊松回归和负二项回归等模型。但是当数据存在过度分散情况时，应该采用负二项回归而非泊松回归模型。为研究自引现象对文献被他引数量的影响，本文使用面板数据，将每篇文献的被他引数量(即被引总数减去自被引数)作为因变量进行回归。自变量以自引用数量、自被引数为核心被解释变量，将文献的引用数量、文献出版年份、作者平均h指数、出版当年期刊影响因子、文献中作者人数作为控制变量。

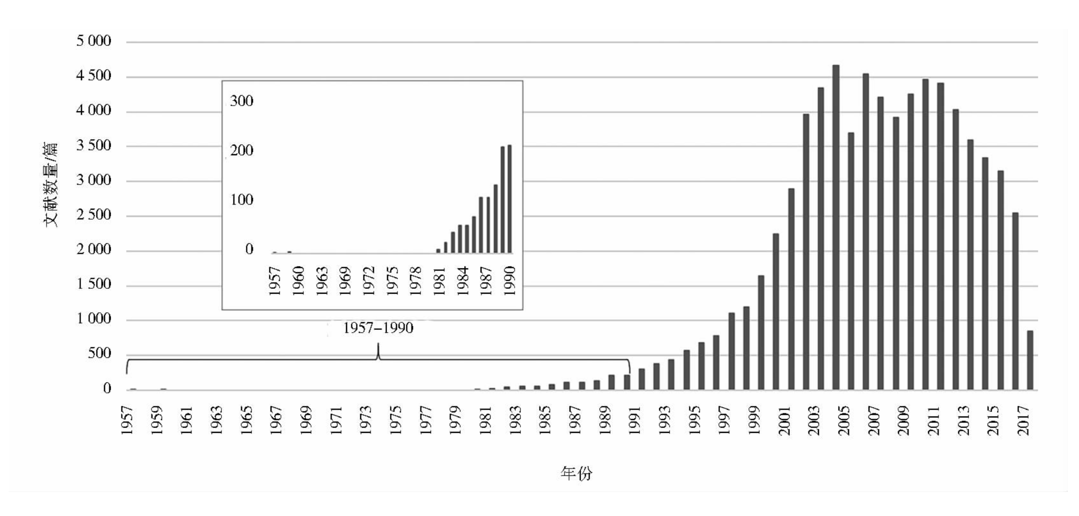


图 1　文献发布时间分布

4　结果分析

4.1　自引总体状况

图 2 显示了 1980 – 2017 各年所发表文献平均自引证率。可见,在 2000 – 2017 年期间,平均自引证率比较稳定,大致在 4% – 6%。2000 – 2017 年,虽然文献自引证数总体上升(2000 年出版文献平均自引 0.13 次,2017 年出版文献平均自引 0.59 次),但由于文献引用数量也在总体上升,因此自引证率变化不大。

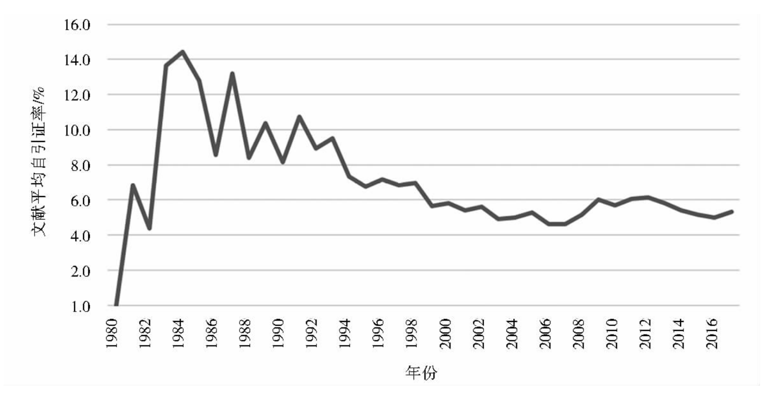


图 2　1980 – 2017 各年所发表文献平均自引证率(不含零施引文献)

数据中 73 160 篇文献一共引用了 430 505 次,其中 4.50% 属于作者自引;这些文献一共被引 811 301 次,其中2.97% 属于作者自被引。有 12 782 篇(17.47%)文献发生了作者自引,有 13 698 篇(18.72%)文献被作者自引。

数据中的 38 556 位作者平均自引证率为 1.82%,作者平均自被引率为 2.05%。作者自引证率和自被引率的分布大致相似,绝大多数作者的自引证率或自被引率都在[0,10%),有少数作者的自引证率或自被引率在[10%,20%),仅有 2.17% 的作者自引证率超过了 20%,2.52% 的作者的自被引率超过了 20%。

4.2　作者自引行为的相关因素分析

4.2.1　作者自被引与文献出版后时间的关系

文献作者自被引平均发生在文献发表后的 2.65 年(当年的自被引视为发表后第 0 年),相比之下,文献的被引用平均发生在发表后 4.47 年。由图 3(a)可知,文献在发表后第一年自被引数达到峰值,而在发表后 2 – 20 年自被引数递减,自被引数也比较少。由图 3(b)(c)(d)可见,对于被引量在不同程度的文献,均显示出自被引行为一般在文献出版后更早发生的现象。

4.2.2　合著文献中作者顺序与作者自引的关系

表 2 显示了在有自引证行为文献中不同顺序作者的自引占比平均值,这个比例是指不同顺序作者在文献中的自引证数与该文献总自引证数的比值。在存在自引证行为的文献中,第一作者的自引占该文献所有自引的比例平均为 73.99%。总体来看,文献中排名越靠后的作者,平均自引百分比越低。对于各种合作规模的文献,第一作者的自引占比平均值在所有作者中都最高。但当作者数从 1 增加到 8(含)以上时,随着作者数量的增加,第一作者的自引占比平均值不断下降。

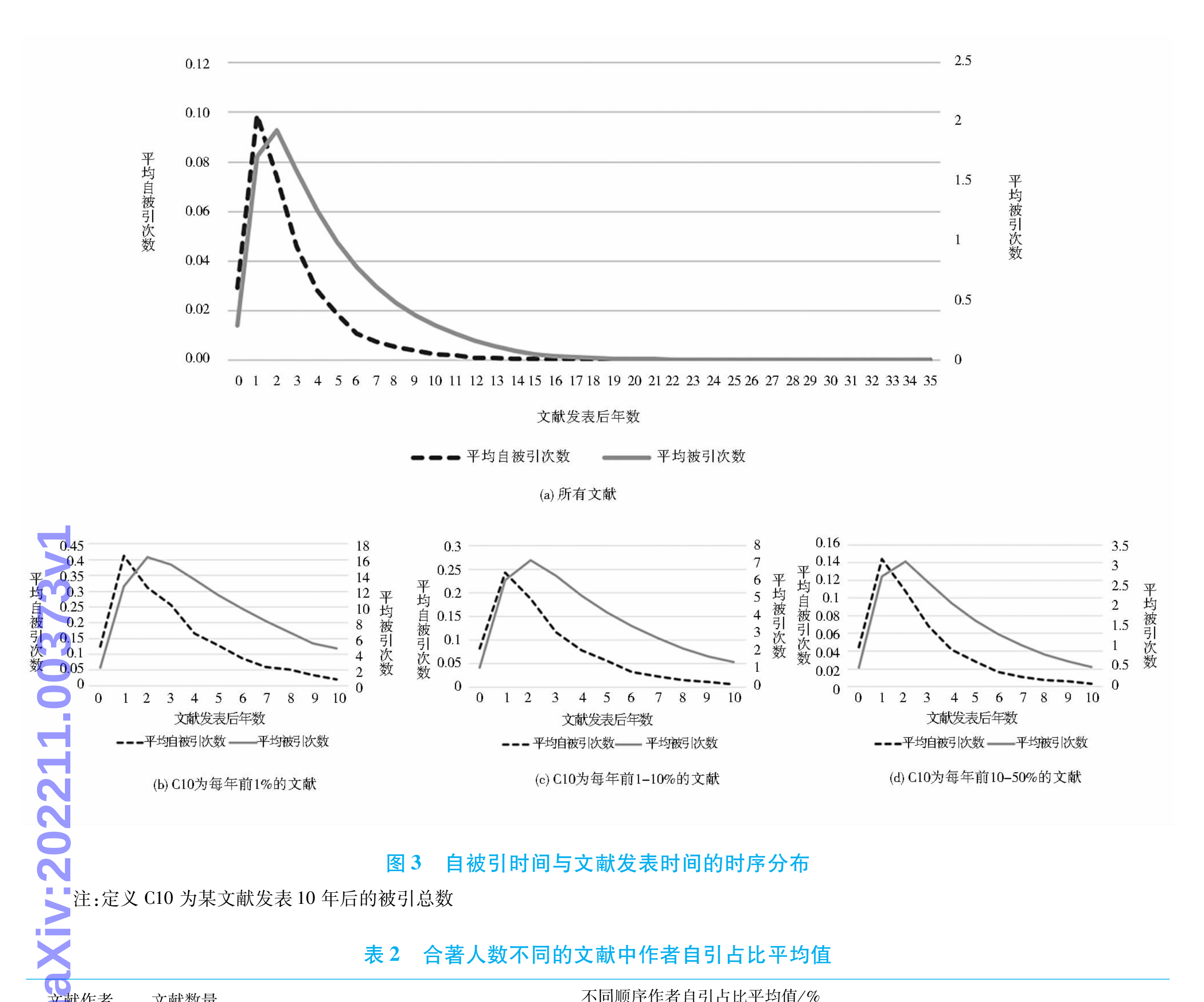


图 3　自被引时间与文献发表时间的时序分布

注:定义 C10 为某文献发表 10 年后的被引总数

表 2　合著人数不同的文献中作者自引占比平均值

文献作者数量/人	文献数量/篇	不同顺序作者自引占比平均值/%							
		1st	2nd	3rd	4th	5th	6th	7th	>7th
1	5 289	100.00							
2	4 030	61.20	38.80						
3	2 255	51.27	27.59	21.14					
4	787	45.99	26.99	14.09	12.93				
5	242	48.02	23.35	10.46	8.81	9.36			
6	96	41.76	19.99	10.59	8.95	7.27	11.44		
7	48	36.35	20.41	8.19	14.84	8.62	4.35	7.25	
>7	35	30.29	10.88	8.62	3.81	11.41	3.93	8.60	22.46
加权平均值/%	73.99	33.20	18.19	11.60	8.97	8.07	7.82	22.46	

4.2.3　作者自引与生产力、学术生涯长度、合作情况的关系

由表 3 可知,作者的自引证数与发表数量、学术生涯长度、合作人数/合作人次有不同程度的正相关性。类似的,作者自被引数也与上述因素存在弱或中度相关性。因此,可以认为在一般情况下,发表文献数量越多、合作过的总人次越多、学术生涯越长、h 指数越高的作者,他们的自引证数或自被引数也较高。另外,这种正相关性是有条件的,当将作者按论文发表数量或 h 指数分组后计算变量相关性发现,对于有一定生产力或影响力的作者群体,能够更容易观测到自引证数/自被引数与合作情况、引用数量、被引数的正相关性,如表 4、表 5 结果所示:

李昕雨，雷佳琪，步 一.我国图书情报学作者自引行为研究初探[J].图书情报工作,2022,66(20):162－171.

表 3　作者自引情况与其他因素的 Spearman 相关性分析

Spearman 相关系数	自引证数	自被引数	自引证率	自被引率
合作人次	.331 **	.347 **	.304 **	.312 **
合作人数	.304 **	.317 **	.277 **	.283 **
篇均合作人次	.042 **	.060 **	.039 **	.059 **
篇均合作人数	－.055 **	－.039 **	－.050 **	－.030 **
发表文献数量	.477 **	.484 **	.445 **	.438 **
引用数量	.405 **	.429 **	.368 **	.392 **
被引数	.375 **	.404 **	.346 **	.351 **
被他引数	.362 **	.372 **	.333 **	.316 **
h 指数	.465 **	.479 **	.432 **	.428 **
剔除自引后 h 指数	.445 **	.443 **	.412 **	.387 **
学术生涯长度	.441 **	.442 **	.413 **	.401 **

注：**. 在 0.01 级别(双尾),相关性显著

表 4　按作者发表文献数量分组的 Spearman 相关性分析

Spearman 相关系数	相关性(分组:发表文献数量 ∈[0,3])				相关性(分组:发表文献数量＞3)			
	自引证数	自被引数	自引用率	自被引率	自引证数	自被引数	自引用率	自被引率
合作人次	.093 **	.114 **	.090 **	.110 **	.389 **	.405 **	.248 **	.246 **
合作人数	.075 **	.094 **	.073 **	.091 **	.347 **	.355 **	.211 **	.203 **
篇均合作人次	.025 **	.047 **	.024 **	.046 **	.101 **	.127 **	.063 **	.107 **
篇均合作人数	－0.000	.018 **	－0.000	.019 **	－0.011	0.007	－0.022	0.017
发表数量	.216 **	.210 **	.209 **	.201 **	.533 **	.524 **	.357 **	.298 **
引用数量	.136 **	.165 **	.126 **	.161 **	.469 **	.494 **	.242 **	.313 **
被引数量	.082 **	.130 **	.079 **	.114 **	.412 **	.418 **	.266 **	.121 **
被他引数	.063 **	.073 **	.060 **	.056 **	.386 **	.384 **	.240 **	.083 **
h 指数	.174 **	.186 **	.168 **	.175 **	.471 **	.476 **	.313 **	.211 **
剔除自引后 h 指数	.137 **	.110 **	.132 **	.097 **	.431 **	.425 **	.273 **	.153 **
学术生涯长度	.207 **	.196 **	.201 **	.189 **	.270 **	.244 **	.176 **	.098 **

注：**. 在 0.01 级别(双尾),相关性显著

表 5　按作者 h 指数分组的 Spearman 相关性分析

Spearman 相关系数	相关性(分组:h 指数 ∈[0,2])				相关性(分组:h 指数＞2)			
	自引证数	自被引数	自引用率	自被引率	自引证数	自被引数	自引用率	自被引率
合作人次	.090 **	.113 **	.087 **	.109 **	.411 **	.421 **	.283 **	.282 **
合作人数	.070 **	.092 **	.068 **	.089 **	.372 **	.377 **	.249 **	.243 **
篇均合作人次	.029 **	.050 **	.028 **	.049 **	.106 **	.130 **	.077 **	.115 **
篇均合作人数	0.003	.020 **	0.003	.021 **	－0.003	0.015	－0.009	.026 *
发表数量	.202 **	.206 **	.195 **	.196 **	.552 **	.538 **	.393 **	.349 **
引用数量	.127 **	.162 **	.117 **	.158 **	.494 **	.513 **	.292 **	.356 **
被引数量	.087 **	.138 **	.084 **	.122 **	.426 **	.419 **	.295 **	.160 **
被他引数	.071 **	.086 **	.068 **	.069 **	.399 **	.384 **	.269 **	.120 **
h 指数	.171 **	.191 **	.165 **	.178 **	.503 **	.497 **	.354 **	.274 **
剔除自引后 h 指数	.142 **	.125 **	.137 **	.110 **	.457 **	.436 **	.308 **	.201 **
学术生涯长度	.192 **	.190 **	.187 **	.181 **	.321 **	.296 **	.225 **	.165 **

注：**. 在 0.01 级别(双尾),相关性显著；*. 在 0.05 级别(双尾),相关性显著

4.3 作者自引对 h 指数的影响

数据显示，包含自引时，作者 h 指数平均为 2.074；而剔除自引后，h 指数平均为 2.028。剔除自引使得作者的 h 指数下降了 0 到 4 不等，作者 h 指数平均下降了 0.045，平均下降比例为 1.83%。通过 Wilcoxon 符号秩检验发现，含自引的 h 指数与不含自引的 h 指数之间有显著差异（p = 0.000）。

为研究图书情报学领域中，影响力较大、知名作者的自引对他们的 h 指数及其排名的影响，本文计算出被引数最高的前 20 位作者 h 指数和被引数在剔除自引后的情况，剔除自引后，他们的 h 指数下降幅度在 0 至6.06% 范围内，有 11 位作者的 h 指数在剔除自引后并没有发生变动。剔除自引后，他们的 h 指数平均下降了 1.77%。

因此，对于图书情报学领域影响力较大的作者群体，是否剔除自引对 h 指数的数值和排名有一定程度的影响，但总体影响较小。

4.4 文献的作者自引对文献被他引数量的影响

由于被他引数量属于计数数据，并且在一定程度上呈现过度分散的特点，因此采用 Stata 的 NB2 负二项回归，使用稳健标准误。回归结果显示，负二项回归 alpha 值的 95% 置信区间为（0.90，0.92），应拒绝 alpha = 0 的原假设，说明数据存在明显的过度离散现象，采用负二项回归比泊松回归更加合适。

回归结果见表 6，自变量中除了“作者人数”，其他回归系数均显著。表 6 回归结果的发生率比值（incidence-rate ratios，IRR）形式解释，IRR 表示解释变量增加 1 时，被解释变量的新的发生率与旧的发生率的比值的平均值。由表 6 可见在其他变量相似的条件下，当文献的自引用数量增加 1 时，被他引的数量平均会降低 9.86%。当文献自被引的数量增加 1 时，其被他引的数量平均会增加 12.82%。由此可见，当文献的作者影响力、所发表的期刊影响力等条件相似的情况下，文献被作者自引时，会增加其可见性，从而增加其被他人引用的可能性。相反，在其他条件相似的情况下，一篇文献的作者自引用的增加，可能会因为文献质量相对较低、受到偏见等原因，获得更少的被引。

表 6 负二项回归结果

Dispersion = mean
Log pseudolikelihood = −242 927.35

Number of obs = 73,160
Wald chi2(7) = 7 150.97
Prob > chi2 = 0.000 0
Pseudo R^2 = 0.028 7

文献被他引数量	Coef.	Robust Std. Err	z	P > |z|	IRR
自引证数量	−0.103 8	0.008 0	−12.92	0.000	0.901 4
自被引数	0.120 6	0.006 4	18.81	0.000	1.128 2
引用数量	0.002 1	0.001 0	2.1	0.036	1.002 1
发表年份	−0.039 6	0.001 1	−36.45	0.000	0.961 1
平均 h 指数	0.047 1	0.001 0	44.98	0.000	1.048 2
当年期刊影响因子	0.285 1	0.007 0	41.01	0.000	1.329 9
作者人数	0.014 7	0.005 8	2.54	0.011	1.52E+35
_cons	81.006 6	2.176 4	37.22	0.000	1.52

5 结论

从分析结果来看，本文的结论可以总结如下：

5.1 中国图书情报学领域的自引行为总体上较少

通过对 1955 – 2017 年间的 73 160 篇文献的分析，得出我国图书情报学领域作者平均自引证率为 1.82%，平均自被引率为 2.05%。所有引用中 4.50% 属于作者自引，所有被引中 2.97% 属于作者自被引，所有文献当中 17.47% 的文献发生了作者自引，18.72% 的文献被作者自引。与以往研究结论[10,14 – 15]相比，可以发现我国图书情报学领域的自引水平相对较低，例如，H. Snyder 和 S. Bonzi 基于 1980 – 1989 年来自人文社科和自然科学多个领域的期刊文献数据的研究发现总体 9% 的引用为自引，其中社会科学领域中约 6% 的引文是自引[16]，尽管自引水平因学科、时间而异，但可以认为我国图书情报学领域自引水平总体上处于合理范围。

5.2 作者自引频次与其发表文献数量、合作总人次、学术生涯长度有不同程度的正相关

这些现象很可能都是因为作者自引频次与作者生产力之间的强正相关性。当作者生产力越高,发表的文献越多,能够自引的机会也就越多,同时也可能因研究之间的延续性和相关性而产生了更强的自引。因此,不能以自引的绝对频次来判断作者是过度自引还是适度自引。

5.3 合著文献中作者顺序与作者自引相关

对于各种合作规模的文献,第一作者在文献中的自引用在该文献的所有自引用中的比重平均值都是最高,这表明第一作者可能更具有科研的积累优势,在研究中承担比较重要的工作[20],而当合作规模扩大时,第一作者的自引比例则有所下降,这与合作程度提高下的研究工作的分担有关。因此,合著文献中自引比重或能在一定程度上说明各作者对文献的贡献程度。

5.4 文献被作者自引与该文献出版后时间相关

文献的自被引平均发生在文献发表后的 2.65 年(当年的自被引视为发表后 0 年)。总体上看,文献在发表后第 2 年自被引数达到高峰,而这以后的自被引数逐渐降低。因此,从文献出版到文献自被引经历的时间间隔总体较短的这一现象,可以推测我国图书情报学领域自引行为较大程度上与研究之间的继承性和相关性有关。

5.5 作者自引对 h 指数数值以及排名有一定影响,但总体影响较小

对于数据中所有作者而言,剔除自引后作者 h 指数平均下降比例为 1.83%。对于中国图书情报学领域被引数排名前 20 的作者而言,剔除自引使其 h 指数平均下降了 1.77%,其中部分学者的 h 指数排名有小幅变动。因此,无论是对我国图书情报学领域的高影响力作者群体,还是范围更大的普通作者群体,可以认为在该领域自引对 h 指数的影响比较小。

5.6 对于文献而言,作者自引用以及自被引会对文献的被他引频次造成影响

在文献作者数量、所发表期刊当年影响因子、作者平均 h 指数、发表时间等因素受控制的条件下,当文献的自引用数量增加 1 时,被他引的数量平均会降低 9.86%;当文献自被引数增加 1 时,其被他引的数量平均会增加 12.82%。由此推测,文献的自引用数量增加时,由于自引相比其他引用的动机更为多样,自引为不当引用的概率也就更高,因此文献整体质量也有可能受不当自引影响而降低,导致其被他引数量也降低。而当文献自被引时,该文献的可见性和影响力便会增加,使得其他研究者更容易注意到该文献,因此该文献被他引的数量也更可能增加。

综上所述,在我国图书情报学领域,作者自引的现象并不十分常见,过度自引的现象很可能更为罕见,并且因作者自引对 h 指数产生的影响可能也比较小。作者自引对文献被他引频次的影响是客观存在的,但这很可能也跟文献自身质量水平相关。

总的来说,虽然目前已有一些识别过度自引或将自引适度纳入科学评价的方法[34－38],但在科学评价实践中类似的方法并未得到推广,也难以证实哪种方法是比较科学、公允。从本文的研究来看,自引与发表文献数量等众多因素相关,而自引行为对 h 指数及其排名有一定影响,但总体影响较小,因此,更应该反对简单地在科学评价中删除自引的方法。而寻找科学处理自引的方法,还需要基于自引相关规律的后续研究和验证。未来,笔者认为自引的研究和处理应该向论文全文本分析、自引功能判断等方面发展,通过引文功能判断、引文动机判断等智能化识别自引行为的合理性。由于时间与技术等原因,本研究还存在诸多的不足之处,包括数据覆盖不够全面、结论局限在图书情报学领域等问题,在未来的研究中可以进一步改进。

参考文献:

[1] 邱均平. 科学文献自引的统计与分析[J]. 情报学刊,1989(6):16－21.
[2] 李韶红,侯金川. 自引与自引分析[J]. 图书馆,2001(6):39－43.
[3] CLARIVATE. HOPKINSON A. Journal citation reports help: title suppressions[EB/OL]. [2022－05－30]. https://jcr.help.clarivate.com/Content/title-suppressions.htm.
[4] 李运景,侯汉清. 自然科学期刊自引对影响因子的“调控”[J]. 情报学报,2006,25(2):7.
[5] ANDRADE A, GONZÁLEZ-JONTE R, CAMPANARIO J. Journals that increase their impact factor at least fourfold in a few years: the role of journal self-citations[J]. Scientometrics, 2009, 80(2):515－528.
[6] SCHREIBER M. Self-citation corrections for the Hirsch index[J]. Europhysics letters, 2007, 78(3):30002.
[7] GARFIELD E. Journal citation studies XVII: Journal self-citation rates-there's a difference[J]. Essays of an information scientist, 1974, 52(2):192－194.

[8] GARFIELD E. Citation indexing-its theory and application in science, technology, and humanities[M]. New York: John Wiley, 1979:245.
[9] 金铁成. 学术期刊自引率使用乱象及其应对策略[J]. 科技与出版, 2016(11):96－98.
[10] GARFIELD E. Journal citation reports: a bibliometric analysis of reference processed for the 1974 [R/OL]. (1975－01－01) [2022－09－10]. http:// http://garfield. library. upenn. edu/papers/jcr1975introduction. pdf.
[11] GARFIELD E. Is citation analysis a legitimate evaluation tool? [J]. Scientometrics, 1979, 1(4): 359－375.
[12]潘云涛,武夷山. 自引、他引:说不尽的故事[J]. 科技导报,2007(24):85.
[13] WOLFGANG G, BART T, BALÁZS S. A bibliometric approach to the role of author self-citations in scientific communication [J]. Scientometrics, 2004, 59(1): 63－77.
[14] DAVARPANAH M R, AMEL F. Author self-citation pattern in science[J]. Library review, 2009, 58(4):301－309.
[15] AKSNES D W. A macro study of self-citation[J]. Scientometrics, 2003,56(2):235－246.
[16] SNYDER H, BONZI S. Patterns of self-citation across disciplines (1980－1989) [J]. Journal of information science, 1998, 24 (6): 431－435.
[17] KING M M, BERGSTROM C T, CORRELL S J, et al. Men set their own cites high: gender and self-citation across fields and over time[J]. Socius: Sociological research for a dynamic world,2017 (3) [2022－08－27]. https://www. zhangqiaokeyan. com/searchResult. html.
[18] MALINIAK D, POWERS R, WALTER B F. The gender citation gap in international relations[J]. International organization, 2013, 67(4): 889－922.
[19] 刘筱敏. 期刊论文数量与期刊自引关系分析[J]. 中国科技期刊研究,2010,21(2):148－150.
[20] 蒋颖,金碧辉,刘筱敏. 期刊论文的作者合作度与合作作者的自引分析[J]. 图书情报工作, 2000, 44(12): 23－28.
[21] 李建辉,王志魁,徐宏,等. 自引对科技期刊影响因子作用的量化研究[J]. 编辑学报, 2007(2):154－157.
[22] FASSOULAKI A, PARASKEVA A, PAPILAS K, et al. Self-citations in six anaesthesia journals and their significance in determining the impact factor[J]. British journal of anaesthesia, 2000, 84 (2): 266－269.
[23] MIMOUNI M, RATMANSKY M, SACHER Y, et al. Self-citation rate and impact factor in pediatrics[J]. Scientometrics, 2016, 108 (3): 1455－1460.
[24] CAMPANARIO J M, GONZÁLEZ L. Journal self -citations that contribute to the impact factor: documents labeled "editorial material" in journals covered by the science citation index[J]. Scientometrics, 2006, 69(2):365 －386.
[25] 温芳芳. 期刊自引与影响因子关系的分区比较与历时分析——以 JCR 收录的管理学期刊为例[J]. 中国科技期刊研究, 2020, 31(8):110－117.
[26] CAMPANARIO J M, MOLINA A. Surviving bad times: the role of citations, self-citations and numbers of citable items in recovery of the journal impact factor after at least four years of continuous decreases[J]. Scientometrics, 2009, 81(3):859－864.
[27] CAMPANARIO J M. The journal citation reports (SCI edition) with and without journal self-citation[J]. Scientometrics, 2018, 27(2):1699－2407.
[28] GIRI R. Influence of selected factors in journals' citations[J]. Aslib journal of information management, 2019, 71(1):90－104.
[29] XIA X D, WU Y W. Journal self-citation analysis of some Chinese sci-tech periodicals[J]. Serials review, 2011, 37(3):171－173.
[30] FOWLER J H, AKSNES D W. Does self-citation pay? [J]. Scientometrics, 2007, 72(3):427－437.
[31]查颖. h 指数与论文自引——以图书情报领域中国学者为例[J]. 图书馆理论与实践,2008,(6):36－38.
[32] HUANG M H, LIN W Y C. Probing the effect of author self-citations on h index: a case study of environmental engineering[J]. Journal of information science, 2011, 37(5):453－461.
[33]朱强, 何峻, 蔡蓉华. 中文核心期刊要目总览. 第 7 版[M]. 北京: 北京大学出版社, 2015.
[34]俞立平,万晓云,琚春华. 学术期刊影响因子过度自引的修正研究——自然影响因子[J]. 情报理论与实践,2019,42(11):62－68.
[35]刘雪立,周志新,方红玲,等. 2005－2007 年我国医学期刊自引率与过度自引的界定[J]. 中国科技期刊研究,2009,20(4):624－626.
[36] HUMPHREY C, KISELEVA O, SCHLEICHER T. A time-series analysis of the scale of coercive journal self-citation and its effect on impact factors and journal rankings [J]. European accounting review,2019, 28(2):335－369.
[37]金铁成. 采用自被引率与 2 年自被引率检测学术期刊过度自引的比较与分析[J]. 中国科技期刊研究,2016,27(9):949－952.
[38] SZOMSZOR M, PENDLEBURY D A, ADAMS J. How much is too much? the difference between research influence and self-citation excess[J]. Scientometrics, 2020, 123(2): 1119－1147.

作者贡献说明:

李昕雨:实验方法设计、实验数据分析、论文撰写;
雷佳琪:论文审阅修改;
步一:研究思路构思、实验数据准备、文献审阅修改。

李昕雨，雷佳琪，步 一. 我国图书情报学作者自引行为研究初探[J]. 图书情报工作,2022,66(20):162－171.

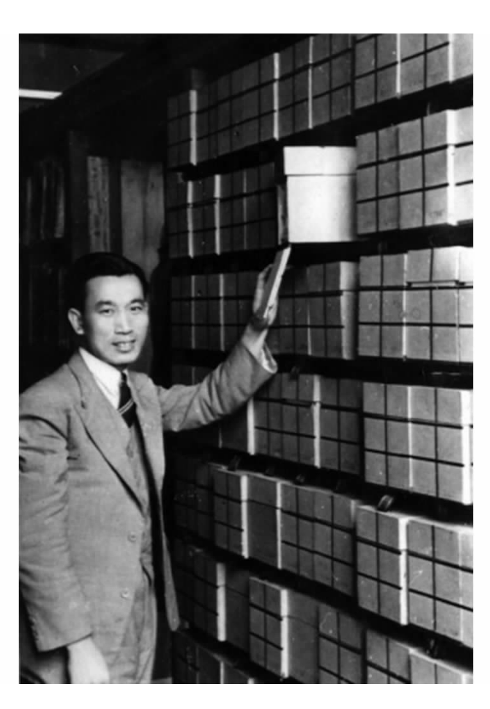
A Preliminary Study on Author Self-Citation Behaviors in Library and Information Science in China

Li Xinyu　Lei Jiaqi　Bu Yi

Department of Information Management, Peking University, Beijing 100871

Abstract: [**Purpose/Significance**] Author self-citations are a common phenomenon and an important factor to be considered in citation analysis and scientific evaluation. However, how to deal with author self-citations in scientific evaluation has always been controversial. This paper expects to provide suggestions for proper treatment of author self-citations by exploring the principles of author self-citations. [**Method/Process**] Based on a thorough survey on previous works related to author self-citations, this paper adopted the journal data of library and information science (Chinese literature) and implemented author- and paper-level analyses of author self-citations by using statistical methods such as regression analysis. [**Result/Conclusion**] Empirical results show that Chinese library and information science researchers tend to have a low rate of author self-citations and that there is a strong positive correlation between the author self-citation frequency and the number of published articles. The paper also observes that author self-citations have limited impact on h-index and that being self-cited may boost the possibility of citations (by others).

Keywords: author self-citation　citation analysis　library and information science　scientific evaluation



1937 年王重民在法国国家图书馆取阅敦煌卷子